# ON A CLASSIFICATION PROBLEM: RANKING AND SELECTION APPROACH *

by

Shanti S. Gupta and Lii-Yuh Leu

Department of Statistics, Purdue University, West Lafayette, IN, USA

Graduate Institute of Statistics, National Central University, Chung-Li, Taiwan, R.O.C.

Technical Report #89-27C

# PURDUE UNIVERSITY

## CENTER FOR STATISTICAL DECISION SCIENCES AND DEPARTMENT OF STATISTICS

**ON A CLASSIFICATION PROBLEM: RANKING AND SELECTION APPROACH ***

by

Shanti S. Gupta
Department of Statistics
Purdue University
West Lafayette, IN, USA

and

Lii-Yuh Leu
Graduate Institute of Statistics
National Central University
Chung-Li, Taiwan, R.O.C.

Technical Report #89-27C



Department of Statistics
Purdue University

August 1989

* This research was supported in part by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grants DMS-8606964, DMS-8702620 at Purdue University. The research of Dr. Lii-Yuh Leu was also supported in part by the National Science Council, Republic of China.

# ON A CLASSIFICATION PROBLEM: RANKING AND SELECTION APPROACH *

by

Shanti S. Gupta and Lii-Yuh Leu
Department of Statistics, Purdue University, West Lafayette, IN, USA
Graduate Institute of Statistics, National Central University, Chung-Li, Taiwan, R.O.C.

## ABSTRACT

This paper deals with a classification problem based on ranking and selection approach. We assume that the populations follow multivariate normal distribution. The corresponding selection problem is to choose the population with the smallest Mahalanobis distance. The subset selection approach is considered throughout this paper. Sometimes the indifference zone approach is also proposed. It should be pointed out that, for the subset selection approach, we need not assume that the individual to be classified belongs to one of the several given categories. The classification procedures depend on whether the parameters $\underline{\mu}_i$ and $\Sigma_i$ are known or unknown.

Key Words and Phrases: classification rules; multivariate normal populations; Mahalanobis distance; ranking and selection; probability of correct classification.



* This research was supported in part by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grants DMS-8606964, DMS-8702620 at Purdue University. The research of Dr. Lii-Yuh Leu was also supported in part by the National Science Council, Republic of China.

## 1. Introduction

The problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several categories on the basis of these measurements. Since Fisher (1936) introduced the linear discriminant function for distinguishing between two multivariate normal distributions with a common covariance matrix, a great deal of work has been done by many authors on this problem. For an extensive bibliography, the reader is referred to Anderson, Das Gupta and Styan (1972), Das Gupta (1973) and Lachenbruch (1975). For a general approach to this problem, Anderson (1984) is a good reference.

In general, for the classification problem, we assume that the individual to be classified belongs to one of the several categories. In a real situation, this assumption may not be appropriate. Thus the problem of selecting the nearest category to the individual based on distance function was considered to cover the above drawback. For the decision theoretic approach see Cacoullos (1965a, 1965b) and Srivastava (1967). Although some intuitive classification procedures have optimality properties based on decision viewpoint, in practice we want to control the probability of misclassification. Using the classical approach, it is difficult to control this probability. Hence an approach based on the concept of ranking and selection was considered by Cacoullos (1973) and A.K. Gupta and Govindarajulu (1973, 1985). Unfortunately, their results are too conservative and very limited.

Let $CC$ stand for a correct classification and $R$ denote a classification procedure, for a given constant $P^*$, $1/k < P^* < 1$, we want to choose a classification procedure R to satisfy the probability requirement (1.1)

$$P(CC|R) \geq P^* \tag{1.1}$$

where $P(CC|R)$ is the probability of a correct classification when the procedure $R$ is used. To make the classification problem more precise, we may ask the problem: can one find a classification procedure satisfying the probability requirement (1.1) and what is the sample size needed?

Let $\pi_i$, $i = 0, 1, \ldots, k$, be $k+1$ populations, we want to classify $\pi_0$ as one of the $\pi_i$, $i =$

$1, \ldots, k$. We assume that $\pi_i \sim N_p\left(\underline{\mu}_i, \Sigma_i\right)$ (the $p$-variate multivariate normal distribution with mean vector $\underline{\mu}_i$ and covariance matrix $\Sigma_i$), $i = 0, 1, \ldots, k$. Based on the Mahalanobis distance between two populations, our problem is related to the problem of selecting the smallest Mahalanobis distance. The subset selection approach is considered throughout this paper. Sometimes the indifference zone approach is also used. The classification procedures depend on whether the parameters $\underline{\mu}_i$ and $\Sigma_i$ are known or unknown.

## 2. Classification procedures when $\underline{\mu}_0$ known, $\underline{\mu}_i$, $i = 1, \ldots, k$, unknown

In this section we assume that $\underline{\mu}_0$ is known and $\underline{\mu}_i$, $i = 1, \ldots, k$, are unknown. The Mahalanobis distance of $\pi_i$ and $\pi_0$ is defined to be $\theta_i = \left(\underline{\mu}_i - \underline{\mu}_0\right)' \Sigma_i^{-1} \left(\underline{\mu}_i - \underline{\mu}_0\right)$, $i = 1, \ldots, k$. Let $\theta_{[1]} \leq \ldots \leq \theta_{[k]}$ denote the ordered values of the $\theta_i$'s, $i = 1, \ldots, k$. Our classification problem may reduce to a selection problem which selects the population corresponding to the parameter $\theta_{[1]}$. We will classify $\pi_0$ as the selected population when the indifference zone approach is used, and classify $\pi_0$ as any one population in the selected subset when the subset selection approach is used.

Let $\underline{X}_{ij}$, $j = 1, \ldots, n$, be a random sample from population $\pi_i$, $\overline{\underline{X}}_i = \frac{1}{n} \sum_{j=1}^{n} \underline{X}_{ij}$ be the sample mean vector, $S_i = \frac{1}{n-1} \sum_{j=1}^{n} \left(\underline{X}_{ij} - \overline{\underline{X}}_i\right) \left(\underline{X}_{ij} - \overline{\underline{X}}_i\right)'$ be the sample covariance matrix for the population $\pi_i$, $i = 1, \ldots, k$, respectively and $S = \frac{1}{k} \sum_{i=1}^{k} S_i$ be the pooled sample covariance matrix of populations $\pi_1, \ldots, \pi_k$. Throughout this paper, we denote $\chi^2_{p;\delta}$ as the noncentral chi-square distribution with degrees of freedom $p$ and noncentrality parameter $\delta$ and $F_{p,q;\delta}$ as the noncentral $F$-distribution with degrees of freedom $p$ and $q$ and noncentrality parameter $\delta$. Also, let $G_p(x;\delta)$ denote the cdf of $\chi^2_{p;\delta}$ and $F_{p,q}(x;\delta)$ the cdf of $F_{p,q;\delta}$. When $\delta = 0$ we simplify these by $\chi^2_p$, $F_{p,q}$, $G_p(x)$ and $F_{p,q}(x)$ respectively.

We will discuss the classification procedures in various situations.

### 2.1. $\Sigma_i$, $i = 1, \ldots, k$, known

#### 2.1.1. Indifference Zone Approach

For the indifference zone approach, we will make the assumptions $\theta_{[1]} = 0$, i.e. $\pi_0$

belongs to one of the $\pi_i$, $i = 1, \ldots, k$, and $\theta_{[2]} \geq \Delta$, where $\Delta$ is a given positive constant. Let $Y_i = n\left(\underline{\overline{X}}_i - \underline{\mu}_0\right)' \Sigma_i^{-1} \left(\underline{\overline{X}}_i - \underline{\mu}_0\right)$, $i = 1, \ldots, k$, then $Y_i \sim \chi^2_{p;n\theta_i}$. Intuitively, we will consider the classification procedure $R_1$ as follows:

$R_1$: Classify $\pi_0$ as $\pi_i$ if and only if $Y_i = \min_{1 \leq j \leq k} Y_j$.

For a given $P^*$, we want to find an appropriate sample size $n$ so that the probability requirement (1.1) is satisfied. The following theorem is useful for this problem.

Theorem 2.1. $\inf P(CC|R_1) = \int_0^\infty [1 - G_p(x; n\Delta)]^{k-1} dG_p(x)$. (2.1)

Proof. Let $Y_{(i)}$ denote the statistic corresponding to the parameter $\theta_{[i]}$. Then

$$\begin{aligned} P(CC|R_1) &= P\left\{Y_{(1)} \leq Y_{(j)},\ j = 2, \ldots, k\right\} \\ &= \int_0^\infty \prod_{j=2}^{k} \left[1 - G_p\left(x; n\theta_{[j]}\right)\right] dG_p(x) \\ &\geq \int_0^\infty [1 - G_p(x; n\Delta)]^{k-1} dG_p(x) \qquad (2.2) \end{aligned}$$

The inequality (2.2) holds, since $G_p(x;\delta)$ has the stochastic increasing property and the equality is attained when $\theta_{[2]} = \ldots = \theta_{[k]} = \Delta$. □

Remark 2.1. Let $d$ be the solution of the equation

$$\int_0^\infty [1 - G_p(x; d)]^{k-1} dG_p(x) = P^* \qquad (2.3)$$

and $n^*$ be the smallest positive integer such that $n\Delta \geq d$. Then $n^*$ is the sample size needed to guarantee the probability requirement (1.1).

### 2.1.2. Subset Selection Approach

For the subset selection approach, Gupta (1966) and Gupta and Studden (1970) have considered the problem for selecting the smallest parameter of $\underline{\mu}_i' \Sigma_i^{-1} \underline{\mu}_i$, $i = 1, \ldots, k$. Following their idea, we consider the classification procedure $R_2$ as follows:

$R_2$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $Y_i \leq c_2 \min_{1 \leq j \leq k} Y_j$, where $c_2 > 1$ is the smallest value such that the probability requirement (1.1) is satisfied.

By applying Theorem 3.1 of Gupta and Studden (1970), we have the following theorem:

Theorem 2.2. $\inf P(CC|R_2) = \int_0^\infty [1 - G_p(x/c_2)]^{k-1} dG_p(x).$ (2.4)

Proof. $P(CC|R_2) = P\left\{Y_{(1)} \le c_2 Y_{(j)},\ j = 2, \ldots, k\right\}$

$$= \int_0^\infty \prod_{j=2}^{k} \left[1 - G_p\left(x/c_2; n\theta_{[j]}\right)\right] dG_p\left(x; n\theta_{[1]}\right)$$

$$\ge \int_0^\infty \left[1 - G_p\left(x/c_2; n\theta_{[1]}\right)\right]^{k-1} dG_p\left(x; n\theta_{[1]}\right) \tag{2.5}$$

$$\ge \int_0^\infty [1 - G_p(x/c_2)]^{k-1} dG_p(x). \tag{2.6}$$

The equality in (2.5) holds when $\theta_{[1]} = \theta_{[2]} = \ldots = \theta_{[k]}$ and the equality in (2.6) holds when $\theta_{[1]} = 0$. □

Remark 2.2. The values $c_2$ satisfying the equation $\int_0^\infty [1 - G_p(x/c_2)]^{k-1} dG_p(x) = P^*$ can be found from Gupta and Sobel (1962) and Armitage and Krishnaiah (1964).

On the other hand, we may consider another classification procedure $R_3$ as follows:

$R_3$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $Y_i \le c_3$, where $c_3$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

For the determination of the value $c_3$, we can use the following theorem:

Theorem 2.3. $\inf P(CC|R_3) = G_p(c_3)$, if $\theta_{[1]} = 0$.

Proof. $P(CC|R_3) = P\left\{Y_{(1)} \le c_3\right\}$

$$= G_p(c_3) \text{ if } \theta_{[1]} = 0.$$ □

The values of $c_3$ can be found from $\chi_p^2$–tables.

Remark 2.3. Note that if we use the procedure $R_3$, the selected subset may be an empty set. This drawback suggests that $\pi_0$ may not belong to one of the $\pi_i$, $i = 1, \ldots, k$.

Remark 2.4. We may consider a testing problem: $H_0: \underline{\mu}_0 = \underline{\mu}_i$, for some $i$, $i = 1, \ldots, k$, vs. $H_1: \underline{\mu}_0 \ne \underline{\mu}_i$, $i = 1, \ldots, k$. For a level $\alpha$ test, the suggested rejection region could be $\min_{1 \le j \le k} Y_j \ge d$, where $d$ is the $c_3$ value determined by the procedure $R_3$ and the associated probability $P^*$ is $1-\alpha$. We note that

$$P_{H_0}\left(\min_{1\le j\le k} Y_j \ge d\right) = 1 - P_{H_0}\left(\min_{1\le j\le k} Y_j < d\right)$$
$$\le 1 - P_{H_0}\left(Y_{(1)} \le d\right) = 1 - P^* = \alpha.$$

## 2.2. $\Sigma_i$, $i=1,\ldots,k$, unknown, not all equal

### 2.2.1. Indifference Zone Approach

If $\Sigma_i$, $i=1,\ldots,k$, are unknown but not all equal, we may estimate $\theta_i$ by $\left(\underline{\overline{X}}_i - \underline{\mu}_0\right)' S_i^{-1} \left(\underline{\overline{X}}_i - \underline{\mu}_0\right)$, $i=1,\ldots,k$. Let $Y_i^* = n\left(\underline{\overline{X}}_i - \underline{\mu}_0\right)' S_i^{-1} \left(\underline{\overline{X}}_i - \underline{\mu}_0\right)$, $i=1,\ldots,k$, and define a classification procedure $R_4$ as follows:

$R_4$: Classify $\pi_0$ as $\pi_i$ if and only if $Y_i^* = \min_{1\le j\le k} Y_j^*$.

We note that $Y_i^* \sim \frac{(n-1)p}{n-p} F_{p,n-p;n\theta_i}$ and $F_{p,n-p}(x;n\theta_i)$ has the stochastic increasing property. If we assume that $\theta_{[1]}=0$ and $\theta_{[2]}\ge\Delta$, then we have the following theorem:

Theorem 2.4. $\inf P(CC|R_4) = \int_0^\infty [1-F_{p,n-p}(x;n\Delta)]^{k-1} dF_{p,n-p}(x)$. (2.7)

Proof. $P(CC|R_4) = P\left\{\frac{n-p}{(n-1)p} Y_{(1)}^* \le \frac{n-p}{(n-1)p} Y_{(j)}^*,\ j=2,\ldots,k\right\}$

$$= \int_0^\infty \prod_{j=2}^k \left[1 - F_{p,n-p}\left(x; n\theta_{[j]}\right)\right] dF_{p,n-p}(x)$$
$$\ge \int_0^\infty [1-F_{p,n-p}(x;n\Delta)]^{k-1} dF_{p,n-p}(x). \qquad \square$$

Remark 2.5. In order to satisfy the probability requirement (1.1), we should choose the smallest $n$ such that $n\Delta \ge d$ and $\int_0^\infty [1-F_{p,n-p}(x;d)]^{k-1} dF_{p,n-p}(x) = P^*$.

### 2.2.2. Subset Selection Approach

Analogous to Gupta and Studden (1970), we consider the classification procedure $R_5$ as follows:

$R_5$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $Y_i^* \le c_5 \min_{1\le j\le k} Y_j^*$, where $c_5 > 1$ is the smallest constant such that the probability requirement (1.1) is satisfied.

By applying Theorem 3.1 of Gupta and Studden (1970), we have the following theorem:

Theorem 2.5. $\inf P(CC|R_5) = \int_0^\infty [1 - F_{p,n-p}(x/c_5)]^{k-1} dF_{p,n-p}(x)$. (2.8)

Proof. $P(CC|R_5) = P\left\{\frac{n-p}{(n-1)p} Y^*_{(1)} \le c_5 \frac{n-p}{(n-1)p} Y^*_{(j)},\ j = 2, \ldots, k\right\}$

$$= \int_0^\infty \prod_{j=2}^{k} \left[1 - F_{p,n-p}\left(x/c_5; n\theta_{[j]}\right)\right] dF_{p,n-p}\left(x; n\theta_{[1]}\right)$$

$$\ge \int_0^\infty \left[1 - F_{p,n-p}\left(x/c_5; n\theta_{[1]}\right)\right]^{k-1} dF_{p,n-p}\left(x; n\theta_{[1]}\right)$$

$$\ge \int_0^\infty [1 - F_{p,n-p}(x/c_5)]^{k-1} dF_{p,n-p}(x). \qquad \square$$

Remark 2.6. The value of $c_5$ is the solution of the equation

$$\int_0^\infty [1 - F_{p,n-p}(x/c_5)]^{k-1} dF_{p,n-p}(x) = P^*.$$

On the other hand, we may consider a classification procedure $R_6$ as follows:

$R_6$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $Y_i^* \le c_6$, where $c_6$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

For the determination of the $c_6$ values, it is easy to show that

Theorem 2.6. $\inf P(CC|R_6) = F_{p,n-p}\left(\frac{n-p}{(n-1)p} c_6\right)$ if $\theta_{[1]} = 0$.

The values of $c_6$ can be found from the $F_{p,n-p}$-tables.

Remark 2.7. Note that if we use the procedure $R_6$, the selected subset may be an empty set.

Remark 2.8. We may consider a testing problem: $H_0: \underline{\mu}_0 = \underline{\mu}_i$, for some $i$, $i = 1, \ldots, k$, vs. $H_1: \underline{\mu}_0 \ne \underline{\mu}_i$, $i = 1, \ldots, k$. For a level $\alpha$ test, the suggested rejection region could be $\min_{1 \le j \le k} Y_j^* \ge d$, where $d$ is the $c_6$ value determined by the procedure $R_6$ and the associated probability $P^*$ is $1 - \alpha$.

## 2.3. $\Sigma_i = \Sigma$, $i = 1, \ldots, k$, unknown

When $\Sigma_i = \Sigma$, $i = 1, \ldots, k$, are unknown, we estimate $\Sigma$ by $S$ and let $Y_i^{**} = n\left(\underline{\overline{X}}_i - \underline{\mu}_0\right)' S^{-1} \left(\underline{\overline{X}}_i - \underline{\mu}_0\right)$, $i = 1, \ldots, k$. We note that $Y_i^{**} \sim \frac{vp}{v-p+1} F_{p,v-p+1;n\theta_i}$, where $v = k(n-1)$ and $\underline{a}'\Sigma^{-1}\underline{a}/Ch_1(S\Sigma^{-1}) \geq \underline{a}'S^{-1}\underline{a} \geq \underline{a}'\Sigma^{-1}\underline{a}/Ch_p(S\Sigma^{-1})$ for any $p \times 1$ vector $\underline{a}$, where $Ch_i(S\Sigma^{-1})$ is the $i$-th smallest eigenvalue of the matrix $S\Sigma^{-1}$. Let $Z = Ch_1(S\Sigma^{-1})/Ch_p(S\Sigma^{-1})$ and $H(z)$ be the cdf of $Z$ ($H$ is independent of $\underline{\mu}_i$, $i = 1,\ldots,k$, and $\Sigma$). This problem was considered by Chattopadhyay (1981) for subset selection approach.

### 2.3.1. Indifference Zone Approach

For the indifference zone approach, we consider the classification procedure $R_7$ as follows:

$R_7$: Classify $\pi_0$ as $\pi_i$ if and only if $Y_i^{**} = \min_{1\leq j\leq k} Y_j^{**}$.

For this procedure, the following result can be used to determine the required sample size to satisfy the probability requirement (1.1).

Theorem 2.7. $\inf P(CC|R_7) \geq \int_0^1 \int_0^\infty [1 - G_p(x/z; n\Delta)]^{k-1} dG_p(x)dH(z)$. (2.9)

Proof. $P(CC|R_7) = P\left\{Y_{(1)}^{**} \leq Y_{(j)}^{**},\ j = 2,\ldots,k\right\}$

$$\geq P\left\{n\left(\underline{X}_{(1)} - \underline{\mu}_0\right)'\Sigma^{-1}\left(\underline{X}_{(1)} - \underline{\mu}_0\right) \leq \frac{Ch_1(S\Sigma^{-1})}{Ch_p(S\Sigma^{-1})} n\left(\underline{X}_{(j)} - \underline{\mu}_0\right)'\Sigma^{-1}\left(\underline{X}_{(j)} - \underline{\mu}_0\right),\right.$$
$$\left. j = 2,\ldots,k\right\}$$
$$= \int_0^1 \int_0^\infty \prod_{j=2}^{k} \left[1 - G_p\left(x/z; n\theta_{[j]}\right)\right] dG_p(x)dH(z)$$
$$\geq \int_0^1 \int_0^\infty [1 - G_p(x/z; n\Delta)]^{k-1} dG_p(x)dH(z). \qquad \square$$

Remark 2.9. Since $Ch_1(S\Sigma^{-1})/Ch_p(S\Sigma^{-1}) \xrightarrow{P} 1$ as $n \to \infty$, we have $P(CC|R_7) \to 1$ as $n \to \infty$.

Remark 2.10. The distribution of $Ch_1(S\Sigma^{-1})/Ch_p(S\Sigma^{-1})$ can be found in Pillai, Al-Ani and Jouris (1969).

Remark 2.11. It is easy to show that $\inf P(CC|R_7) \geq P(Z > \theta_\varepsilon) \int_0^\infty [1 - G_p(x/\theta_\varepsilon; n\Delta)]^{k-1} dG_p(x)$, where $\varepsilon > 0$ and $P(Z \leq \theta_\varepsilon) = \varepsilon$. The equation $P(Z > \theta_\varepsilon) \int_0^\infty [1 - G_p(x/\theta_\varepsilon; n\Delta)]^{k-1} dG_p(x) =$

$P^*$ can be used to determine the required sample size.

2.3.2. Subset Selection Approach

For the subset selection approach we refer to Chattopadhyay (1981). We consider the classification procedure $R_8$ as follows:

$R_8$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $Y_i^{**} \leq c_8 \min_{1\leq j\leq k} Y_j^{**}$, where $c_8 > 1$ is the smallest constant such that the probability requirement (1.1) is satisfied.

Analogous to the proof of Theorem 2.7, we have the following result:

Theorem 2.8. $\inf P(CC|R_8) \geq \int_0^1 \int_0^\infty [1 - G_p(x/c_8 z)]^{k-1} dG_p(x) dH(z).$ (2.10)

Remark 2.12. $\int_0^1 \int_0^\infty [1 - G_p(x/c_8 z)]^{k-1} dG_p(x) dH(z) \to 1$ as $n \to \infty$.

Remark 2.13. It is easy to show that

$$\inf P(CC|R_8) \geq P(Z > \theta_\epsilon) \int_0^\infty [1 - G_p(x/c_8\theta_\epsilon)]^{k-1} dG_p(x).$$

On the other hand, we may consider an easier classification procedure $R_9$ defined as follows:

$R_9$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $Y_i^{**} \leq c_9$, where $c_9$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

It is easy to show that

Theorem 2.9. $\inf P(CC|R_9) = F_{p,v-p+1}\left(\frac{v-p+1}{vp} c_9\right)$ if $\theta_{[1]} = 0$.

Remark 2.14. We can consider the testing problem: $H_0: \underline{\mu}_0 = \underline{\mu}_i$, for some $i$, $i = 1, \ldots, k$, vs. $H_1: \underline{\mu}_0 \neq \underline{\mu}_i$, $i = 1, \ldots, k$. The reject region is $\min_{1\leq j\leq k} Y_j^{**} \geq d$.

## 3. Classification procedures when $\underline{\mu}_0$ unknown, $\underline{\mu}_i$, $i = 1, \ldots, k$, known

In the case that $\underline{\mu}_0$ is unknown and $\underline{\mu}_i$, $i = 1, \ldots, k$, are known. Let $\underline{X}_{01}, \ldots, \underline{X}_{0n}$ be a random sample from $\pi_0$ and $\overline{\underline{X}}_0 = \frac{1}{n} \sum_{j=1}^{n} \underline{X}_{ij}$, $S_0 = \frac{1}{n-1} \sum_{j=1}^{n} \left(\underline{X}_{0j} - \overline{\underline{X}}_0\right)\left(\underline{X}_{0j} - \overline{\underline{X}}_0\right)'$. The Mahalanobis distance between populations $\pi_0$ and $\pi_i$ is defined to be $\lambda_i = \left(\underline{\mu}_i - \underline{\mu}_0\right)' \Sigma_0^{-1} \left(\underline{\mu}_i - \underline{\mu}_0\right)$. We will discuss the classification procedures in various situations.

### 3.1. $\Sigma_0$ known

#### 3.1.1. Indifference Zone Approach

For the indifference zone approach, we assume that $\lambda_{[2]} - \lambda_{[1]} \geq \Delta$. We define $Z_i = \left(\overline{X}_0 - \underline{\mu}_i\right)' \Sigma_0^{-1} \left(\overline{X}_0 - \underline{\mu}_i\right)$, $i = 1, \ldots, k$. Intuitively, we may consider the classification procedure $R_{10}$ as follows:

$R_{10}$: Classify $\pi_0$ as $\pi_i$ if and only if $Z_i = \min\limits_{1 \leq j \leq k} Z_j$.

For the procedure $R_{10}$, we note that

$$Z_i \leq Z_j,\ j = 1, \ldots, k,\ j \neq i \text{ iff } 2\left(\underline{\mu}_j - \underline{\mu}_i\right)' \Sigma_0^{-1} \left(\overline{X}_0 - \underline{\mu}_0\right) \leq \lambda_j - \lambda_i.$$

Thus we have the following result:

Theorem 3.1. $\inf P(CC|R_{10}) \geq 1 - (k-1)\Phi\left(-\frac{\sqrt{n}\Delta}{2\delta}\right)$, (3.1)

where $\Phi$ is the cdf of the standard normal and $\delta^2 = \max\limits_{1 \leq i \leq k} \delta_i^2$, $\delta_i^2 = \max\limits_{\substack{1 \leq j \leq k \\ j \neq i}} \left(\underline{\mu}_j - \underline{\mu}_i\right)' \Sigma_0^{-1} \left(\underline{\mu}_j - \underline{\mu}_i\right)$.

Proof. $P(CC|R_{10}) = P\left\{Z_{(1)} \leq Z_{(j)},\ j = 2, \ldots, k\right\}$

$$\begin{aligned}
&= P\left\{2\left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)' \Sigma_0^{-1} \left(\overline{X}_0 - \underline{\mu}_0\right) \leq \lambda_{[j]} - \lambda_{[1]},\ j = 2, \ldots, k\right\} \\
&\geq P\left\{2\left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)' \Sigma_0^{-1} \left(\overline{X}_0 - \underline{\mu}_0\right) \leq \Delta,\ j = 2, \ldots, k\right\} \\
&\geq 1 - \sum_{j=2}^{k} P\left\{2\left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)' \Sigma_0^{-1} \left(\overline{X}_0 - \underline{\mu}_0\right) > \Delta\right\} \\
&= 1 - \sum_{j=2}^{k} \Phi\left(\frac{-\sqrt{n}\Delta}{2\left[\left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)' \Sigma_0^{-1} \left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)\right]^{1/2}}\right) \\
&\geq 1 - (k-1)\Phi\left(\frac{-\sqrt{n}\Delta}{2\delta}\right). \qquad (3.2)
\end{aligned}$$

The inequality (3.2) holds since $\delta^2 \geq \left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)' \Sigma_0^{-1} \left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)$. □

Remark 3.1. As $n \to \infty$, $\Phi\left(\frac{-\sqrt{n}\Delta}{2\delta}\right) \to 0$, hence $P(CC|R_{10}) \to 1$. For given $P^*$, we can find $n$ such that $\inf P(CC|R_{10}) \geq P^*$.

3.1.2. Subset Selection Approach

For the subset selection approach, we consider a classification procedure $R_{11}$ as follows:

$R_{11}$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $Z_i \leq \min\limits_{1\leq j\leq k} Z_j + c_{11}$, where $c_{11}$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

Analogous to the proof of Theorem 3.1, we have the following result:

Theorem 3.2. $\inf P(CC|R_{11}) \geq 1-(k-1)\Phi\left(\frac{-\sqrt{n}c_{11}}{2\delta}\right)$. (3.3)

Proof. $P(CC|R_{11}) = P\left\{Z_{(1)} \leq Z_{(j)} + c_{11},\ j=2,\ldots,k\right\}$

$$= P\left\{2\left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)' \Sigma_0^{-1}\left(\overline{\underline{X}}_0 - \underline{\mu}_0\right) \leq \lambda_{[j]} - \lambda_{[1]} + c_{11},\ j=2,\ldots,k\right\}$$

$$\geq P\left\{2\left(\underline{\mu}_{(j)} - \underline{\mu}_{(1)}\right)' \Sigma_0^{-1}\left(\overline{\underline{X}}_0 - \underline{\mu}_0\right) \leq c_{11},\ j=2,\ldots,k\right\}$$

$$\geq 1-(k-1)\Phi\left(\frac{-\sqrt{n}c_{11}}{2\delta}\right). \qquad \square$$

On the other hand, we suggest another classification procedure $R_{12}$ as follows:

$R_{12}$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $nZ_i \leq c_{12}$, where $c_{12}$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

For this procedure, it is easy to show that

Theorem 3.3. $\inf P(CC|R_{12}) = G_p(c_{12})$ if $\lambda_{[1]} = 0$.

Remark 3.2. We may consider a testing problem: $H_0\colon \underline{\mu}_0 = \underline{\mu}_i$, for some $i$, $i=1,\ldots,k$, vs. $H_1\colon \underline{\mu}_0 \neq \underline{\mu}_i$, $i=1,\ldots,k$. The suggested reject region is $\min\limits_{1\leq j\leq k} Z_j \geq d$.

3.2. $\Sigma_0$ unknown

When $\Sigma_0$ is also unknown, we estimate it by $S_0$, and let $Z_i^* = n\left(\overline{\underline{X}}_0 - \underline{\mu}_i\right)' S_0^{-1}\left(\overline{\underline{X}}_0 - \underline{\mu}_i\right)$, $i=1,\ldots,k$. For the indifference zone approach, we consider the classification procedure $R_{13}$ as follows:

$R_{13}$: Classify $\pi_0$ as $\pi_i$ if and only if $Z_i^* = \min\limits_{1\leq j\leq k} Z_j^*$.

Analogous to the proof of Theorem 2.7, and Theorem 3.1, for large sample, we have the following result:

Theorem 3.4. $\inf P(CC|R_{13}) \geq 1-(k-1)\Phi\left(\frac{-\sqrt{n}\Delta}{2\delta}\right)$, if $n$ is large enough.

Proof. $P(CC|R_{13}) = P\left\{Z^*_{(1)} \leq Z^*_{(j)},\ j=2,\ldots,k\right\}$

$$\geq P\left\{n\left(\underline{\overline{X}}_0-\underline{\mu}_{(1)}\right)'\Sigma_0^{-1}\left(\underline{\overline{X}}_0-\underline{\mu}_{(1)}\right) \leq \frac{Ch_1(S_0\Sigma_0^{-1})}{Ch_p(S_0\Sigma_0^{-1})}n\left(\underline{\overline{X}}_0-\underline{\mu}_{(j)}\right)'\Sigma_0^{-1}\left(\underline{\overline{X}}_0-\underline{\mu}_{(j)}\right),\ j=2,\ldots,k\right\}$$

$$\approx P\left\{n\left(\underline{\overline{X}}_0-\underline{\mu}_{(1)}\right)'\Sigma_0^{-1}\left(\underline{\overline{X}}_0-\underline{\mu}_{(1)}\right) \leq n\left(\underline{\overline{X}}_0-\underline{\mu}_{(j)}\right)'\Sigma_0^{-1}\left(\underline{\overline{X}}_0-\underline{\mu}_{(j)}\right),\ j=2,\ldots,k\right\}$$

$$\geq 1-(k-1)\Phi\left(\frac{-\sqrt{n}\Delta}{2\delta}\right),\ \text{if } n \text{ is large enough.} \qquad \square$$

For the subset selection approach, we consider the classification procedure $R_{14}$ as follows:

$R_{14}$: Classify $\pi_0$ as one of the $\pi_i$'s for which $Z^*_i \leq \min_{1\leq j\leq k} Z^*_j + c_{14}$, where $c_{14}$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

For large sample, it is easy to show that

Theorem 3.5. $\inf P(CC|R_{14}) \geq 1-(k-1)\Phi\left(\frac{-\sqrt{n}c_{14}}{2\delta}\right)$.

On the other hand, a simple procedure $R_{15}$ can be defined as follows:

$R_{15}$: Classify $\pi_0$ as one of the $\pi_i$'s for which $Z^*_i \leq c_{15}$, where $c_{15}$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

Since $Z^*_i \sim \frac{(n-1)p}{n-p}F_{p,n-p;n\lambda_i}$, it is easy to show that

Theorem 3.6. $\inf P(CC|R_{15}) = F_{p,n-p}\left(\frac{n-p}{(n-1)p}c_{15}\right)$ if $\lambda_{[1]}=0$.

## 4. Classification procedures when $\underline{\mu}_i$, $i=0,1,\ldots,k$, unknown

When $\underline{\mu}_i$, $i=0,1,\ldots,k$, are unknown, we use $\theta_i$ defined in Section 2 as a measurement of distance between populations $\pi_0$ and $\pi_i$, $i=1,\ldots,k$, respectively. Let $\underline{X}_{ij}$, $j=1,\ldots,n$, be a random sample from population $\pi_i$, $i = 0, 1, \ldots, k$, $\underline{\overline{X}}_i = \frac{1}{n}\sum_{j=1}^{n} \underline{X}_{ij}$ be the sample mean vector and $S_i = \frac{1}{n-1}\sum_{j=1}^{n}\left(\underline{X}_{ij} - \underline{\overline{X}}_i\right)\left(\underline{X}_{ij} - \underline{\overline{X}}_i\right)'$ be the sample covariance matrix within the population $\pi_i$ and $S^* = \frac{1}{k+1}\sum_{i=0}^{k} S_i$ is the pooled sample covariance matrix. We will discuss the classification procedures in various situations.

4.1. $\Sigma_i$, $i = 1, \ldots, k$, known

When $\Sigma_i$, $i = 1, \ldots, k$, are known. We define $\cup_i = n\left(\underline{\overline{X}}_0 - \underline{\overline{X}}_i\right)' \Sigma_i^{-1}\left(\underline{\overline{X}}_0 - \underline{\overline{X}}_i\right)$, $i = 1, \ldots, k$. For this case, we use the subset selection approach and consider a classification procedure $R_{16}$ as follows:

$R_{16}$: Classify $\pi_0$ as one of the $\pi_i$'s for which $\cup_i \leq c_{16} \min\limits_{1\leq j\leq k} \cup_j$, where $c_{16} > 1$ is the smallest constant such that the probability requirement (1.1) is satisfied.

For this procedure, we have the following result:

Theorem 4.1. $\inf P(CC|R_{16}) \geq \int_0^\infty [1 - G_p(2x/c_{16})]^{k-1} dG_p(x)$. (4.1)

Proof. Let $\cup_{(j)} = n\left(\underline{\overline{X}}_0 - \underline{\overline{X}}_{(j)}\right)' \Sigma_{(j)}^{-1}\left(\underline{\overline{X}}_0 - \underline{\overline{X}}_{(j)}\right)$, $j = 1, \ldots, k$. Given $\underline{\overline{X}}_0 = \underline{x}$, $\cup_{(j)}$, $j = 1, \ldots, k$, are independent and $\cup_{(j)} \sim \chi^2_{p;n\left(\underline{\mu}_{(j)} - \underline{x}\right)' \Sigma_{(j)}^{-1}\left(\underline{\mu}_{(j)} - \underline{x}\right)}$.

$$
\begin{aligned}
P(CC|R_{16}) &= P\left\{\cup_{(1)} \leq c_{16}\cup_{(j)},\ j = 2, \ldots, k\right\} \\
&= \int P\left\{n\left(\underline{\overline{X}}_{(1)} - \underline{x}\right)' \Sigma_{(1)}^{-1}\left(\underline{\overline{X}}_{(1)} - \underline{x}\right) \leq c_{16} n\left(\underline{\overline{X}}_{(j)} - \underline{x}\right)' \Sigma_{(j)}^{-1}\left(\underline{\overline{X}}_{(j)} - \underline{x}\right),\ j = 2, \ldots, k\right\} dF(\underline{x}) \\
&= \int \prod_{j=2}^{k}\left[1 - G_p\left(y/c_{16}; n\left(\underline{\mu}_{(j)} - \underline{x}\right)' \Sigma_{(j)}^{-1}\left(\underline{\mu}_{(j)} - \underline{x}\right)\right)\right] dG_p\left(y; n\left(\underline{\mu}_{(1)} - \underline{x}\right)' \Sigma_{(1)}^{-1}\left(\underline{\mu}_{(1)} - \underline{x}\right)\right) dF(\underline{x}) \\
&\geq \int \prod_{j=2}^{k}[1 - G_p(y/c_{16})] dG_p\left(y; n\left(\underline{\mu}_{(1)} - \underline{x}\right)' \Sigma_{(1)}^{-1}\left(\underline{\mu}_{(1)} - \underline{x}\right)\right) dF(\underline{x}) \\
&= P\left\{\cup_{(1)} \leq c_{16}\cup^*_{(j)},\ j = 2, \ldots, k\right\}
\end{aligned}
$$

where $F$ is the cdf of $\underline{\overline{X}}_0$, $\cup^*_{(j)}$, $j = 2, \ldots, k$, and $\cup_{(1)}$ are independent and $\cup^*_{(j)} \sim \chi^2_p$. Now $\cup_{(1)} \sim 2\chi^2_p$. Hence

$$
P(CC|R_{16}) \geq \int_0^\infty [1 - G_p(2y/c_{16})]^{k-1} dG_p(y). \qquad \square
$$

Remark 4.1. If we use the measurement $v_i = \left(\underline{\mu}_i - \underline{\mu}_0\right)' (\Sigma_0 + \Sigma_i)^{-1} \left(\underline{\mu}_i - \underline{\mu}_0\right)$. Then we have an easier classification procedure $R_{17}$ defined by

$R_{17}$: Classify $\pi_0$ as one of the $\pi_i$'s for which $\cup_i^* \leq c_{17}$, where $\cup_i^* = n\left(\underline{\overline{X}}_0 - \underline{\overline{X}}_i\right)' (\Sigma_0 + \Sigma_i)^{-1} \left(\underline{\overline{X}}_0 - \underline{\overline{X}}_i\right)$ and $c_{17}$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

It is easy to show that $\inf P(CC|R_{17}) = G_p(c_{17})$ if $v_{[1]} = 0$.

## 4.2. $\Sigma_i$, $i = 1, \ldots, k$, unknown, not all equal

When $\Sigma_i$, $i = 1, \ldots, k$, are unknown and $\Sigma_i$, $i = 1, \ldots, k$, are not all equal. We define $V_i = n\left(\underline{\overline{X}}_0 - \underline{\overline{X}}_i\right)' S_i^{-1} \left(\underline{\overline{X}}_0 - \underline{\overline{X}}_i\right)$, $i = 1, \ldots, k$. A classification procedure $R_{18}$ is defined as follows:

$R_{18}$: Classify $\pi_0$ as one of the $\pi_i$'s for which $V_i \leq c_{18} \min_{1 \leq j \leq k} V_j$, where $c_{18} > 1$ is the smallest constant such that the probability requirement (1.1) is satisfied.

Given $\underline{\overline{X}}_0 = \underline{x}$, we have $V_i \sim \frac{(n-1)p}{n-p} F_{p, n-p; n\left(\underline{\mu}_i - \underline{x}\right)' \Sigma_i^{-1} \left(\underline{\mu}_i - \underline{x}\right)}$. Analogous to the proof of Theorem 4.1, we have the following result:

Theorem 4.2. $\inf P(CC|R_{18}) \geq \int_0^\infty [1 - F_{p,n-p}(2y/c_{18})]^{k-1} dF_{p,n-p}(y).$ (4.2)

Proof. $\int P\left\{V_{(1)} \leq c_{18} V_{(j)},\ j = 2, \ldots, k | \underline{\overline{X}}_0 = \underline{x}\right\} dF(\underline{x})$

$$\geq P\left\{V_{(1)} \leq c_{18} V_{(j)}^*,\ j = 2, \ldots, k\right\}$$

where $V_{(j)}^*$ and $V_{(1)}$ are independent, $V_{(1)} \sim \frac{2(n-1)p}{n-p} F_{p,n-p}$ and $V_{(j)}^* \sim \frac{(n-1)p}{n-p} F_{p,n-p}$, $j = 2, \ldots, k$. Thus

$$P(CC|R_{18}) \geq \int_0^\infty [1 - F_{p,n-p}(2y/c_{18})]^{k-1} dF_{p,n-p}(y).$$

Remark 4.2. We can define an easier classification procedure $R_{19}$ as follows:

$R_{19}$: Classify $\pi_0$ as one of the $\pi_i$'s for which $V_i \leq c_{19}$, where $c_{19}$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

It is easy to show that $\inf P(CC|R_{19}) = F_{p,n-p}\left(\frac{n-p}{2(n-1)p} c_{19}\right)$, if $\theta_{[1]} = 0$.

4.3. $\Sigma_i = \Sigma,\ i = 0, 1, \ldots, k$, unknown

When $\Sigma_i = \Sigma,\ i = 0, 1, \ldots, k$, and $\Sigma$ is unknown, we estimate $\Sigma$ by $S^*$ and define $W_i = n\left(\underline{\overline{X}}_0 - \underline{\overline{X}}_i\right)' S^{*-1}\left(\underline{\overline{X}}_0 - \underline{\overline{X}}_i\right),\ i = 1, \ldots, k$. Then a classification $R_{20}$ can be defined by:

$R_{20}$: Classify $\pi_0$ as any one of the $\pi_i$'s for which $W_i \leq c_{20} \min_{1 \leq j \leq k} W_j$, where $c_{20} > 1$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

Given $\underline{\overline{X}}_0 = \underline{x}$, $W_i \sim \frac{v^* p}{v^* - p + 1} F_{p, v^* - p + 1; n\left(\underline{\mu}_i - \underline{x}\right)' \Sigma^{-1}\left(\underline{\mu}_i - \underline{x}\right)}$, where $v^* = (k+1)(n-1)$. Analogous to the proof of Theorem 4.1 and Theorem 2.7, we have the following result:

Theorem 4.3. $\inf P(CC|R_{20}) \geq \int_0^1 \int_0^\infty [1 - F_{p, v^* - p + 1}(2y/c_{20}z)]^{k-1}$

$$dF_{p, v^* - p + 1}(y) dH^*(z). \tag{4.3}$$

where $H^*(z)$ is the cdf of $Ch_1(S^*\Sigma^{-1})/Ch_p(S^*\Sigma^{-1})$.

Proof. $P\left\{W_{(1)} \leq c_{20} W_{(j)},\ j = 2, \ldots, k | \underline{\overline{X}}_0 = \underline{x}\right\}$

$$= P\left\{n\left(\underline{\overline{X}}_{(1)} - \underline{x}\right)' S^{*-1}\left(\underline{\overline{X}}_{(1)} - \underline{x}\right) \leq c_{20} n\left(\underline{\overline{X}}_{(j)} - \underline{x}\right)' S^{*-1}\left(\underline{\overline{X}}_{(j)} - \underline{x}\right),\ j = 2, \ldots, k\right\}$$

$$\geq P\left\{n\left(\underline{\overline{X}}_{(1)} - \underline{x}\right)' \Sigma^{-1}\left(\underline{\overline{X}}_{(1)} - \underline{x}\right) \leq \frac{Ch_1(S^*\Sigma^{-1})}{Ch_p(S^*\Sigma^{-1})} c_{20} n\left(\underline{\overline{X}}_{(j)} - \underline{x}\right)' \Sigma^{-1}\left(\underline{\overline{X}}_{(j)} - \underline{x}\right),\ j = 2, \ldots, k\right\}$$

$$= \int_0^1 \int_0^\infty \prod_{j=2}^{k}\left[1 - F_{p, v^* - p + 1}\left(y/c_{20}z; n\left(\underline{\mu}_{(j)} - \underline{x}\right)' \Sigma^{-1}\left(\underline{\mu}_{(j)} - \underline{x}\right)\right)\right] dF_{p, v^* - p + 1}$$

$$\left(y; n\left(\underline{\mu}_{(1)} - \underline{x}\right)' \Sigma^{-1}\left(\underline{\mu}_{(1)} - \underline{x}\right)\right) dH^*(z)$$

$$\geq \int_0^1 \int_0^\infty [1 - F_{p, v^* - p + 1}(y/c_{20}z)]^{k-1} dF_{p, v^* - p + 1}\left(y; n\left(\underline{\mu}_{(1)} - \underline{x}\right)' \Sigma^{-1}\left(\underline{\mu}_{(1)} - \underline{x}\right)\right) dH^*(z)$$

$$= P\left\{W_{(1)} \leq c_{20} W^*_{(j)},\ j = 2, \ldots, k | \underline{\overline{X}}_0 = \underline{x}\right\}$$

where $W^*_{(j)} \sim \frac{v^* p}{v^* - p + 1} F_{p, v^* - p + 1}$ is independent of $\underline{\overline{X}}_0$. Therefore

$$P(CC|R_{20}) = P\left\{W_{(1)} \leq c_{20} W_{(j)},\ j = 2, \ldots, k\right\}$$

$$\geq \int_0^1 \int_0^\infty [1 - F_{p, v^* - p + 1}(2y/c_{20}z)]^{k-1} dF_{p, v^* - p + 1}(y) dH^*(z). \qquad \square$$

Remark 4.3. We also have $\inf P(CC|R_{20}) \geq P(Z^* > \theta_\epsilon)$

$$\int_0^\infty \left[1 - F_{p,v^*-p+1}\left(2y/c_{20}\theta_\epsilon\right)\right]^{k-1} dF_{p,v^*-p+1}(y),$$

where $Z^* = Ch_1(S^*\Sigma^{-1})/Ch_p(S^*\Sigma^{-1})$.

Remark 4.4. Since $W_i \sim 2\frac{v^* p}{v^*-p+1} F_{p,v^*-p+1;n\theta_i}$, we can define an easier classification procedure $R_{21}$ as follows:

$R_{21}$: Classify $\pi_0$ as one of the $\pi_i$'s for which $W_i \leq c_{21}$, where $c_{21}$ is the smallest positive constant such that the probability requirement (1.1) is satisfied.

It is easy to show that $\inf P(CC|R_{21}) = F_{p,v^*-p+1}\left(\frac{v^*-p+1}{2v^* p} c_{21}\right)$ if $\theta_{[1]} = 0$.

# BIBLIOGRAPHY


Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, Second Edition, John Wiley & Sons, Inc., New York.

Anderson, T.W., Das Gupta, S. and Styan, G.P.H. (1972), *A Bibliography of Multivariate Statistical Analysis*, Oliver and Boyd, Edinburg.

Armitage, J.V. and Krishnaiah, P.R. (1964), Tables for studentized largest chi–square distribution and their application, ARL 64–188, Air Force Base, Ohio.

Cacoullos, T. (1965a), Comparing Mahalanobis distances I: Comparing distances between $k$ known normal populations and another unknown, *Sankhyá* **A27**, 1–22.

Cacoullos, T. (1965b), Comparing Mahalanobis distances II: Bayes procedures when the mean vectors are unknown, *Sankhyá* **A27**, 23–32.

Cacoullos, T. (1973), Distance, discrimination and error, *Discriminant Analysis and Applications*, Cacoullos, T. ed., Academic Press, Inc., New York, 61–75.

Chattopadhyay, A.K. (1981), Selecting the normal population with largest (smallest) value of Mahalanobis distance from the origin, *Commun. Statist.* **A10(1)**, 31–37.

Das Gupta, S. (1973), Theories and methods in classification: A Review, *Discriminant Analysis and Applications*, Cacoullos, T. ed., Academic Press, Inc., New York, 77–137.

Fisher, R.A. (1936), The use of multiple measurement in taxonomic problems, *Ann. Eugen.* **7**, 179–188.

Gupta, A.K. and Govindarajulu, Z. (1973), Some new classification rules for $c$ univariate normal populations, *Canad. J. Statist.* **1**, 139–157.

Gupta, A.K. and Govindarajulu, Z. (1985), On minimum distance classification rules for $c$ multivariate populations, *Statistica*, anno XLV, n. 1, 101–104.

Gupta, S.S. (1966), On some selection and ranking procedures for multivariate normal populations using distance functions, *Multivariate Analysis*, Krishnaiah, P.R. ed., Academic Press, Inc., New York, 457–475.

Gupta, S.S. and Sobel, M. (1962), On the smallest of several correlated $F$-statistics, *Biometrika* **49**, 509–523.

Gupta, S.S. and Studden, W.J. (1970), On some selection and ranking procedure with applications to multivariate populations, *Essays in Probability and Statistics*, Bose, R.C. ed., Statistical Publishing Society, Calcutta, 327–338.

Lachenbruch, P.A. (1975), *Discriminant Analysis*, Hafner Press, A Division of Macmillan Publishing Co., Inc.

Pillai, K.C.S., Al–Ani, S. and Jouris, G.M. (1969), On the distributions of the ratios of the roots of a covariance matrix and Wilk's Criterion for tests of three hypotheses. *Ann. of Math. Statist.* **40**, 2033–2040.

Srivastava, M.S. (1967), Comparing distances between multivariate populations — the problem of minimum distance. *Ann. of Math. Statist.* **38**, 550–556.

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>Unclassified | 1b. RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3. DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release, distribution unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>Technical Report #89-27C | 5. MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a. NAME OF PERFORMING ORGANIZATION<br>Purdue University<br>6b. OFFICE SYMBOL *(If applicable)* | 7a. NAME OF MONITORING ORGANIZATION |
| 6c. ADDRESS *(City, State, and ZIP Code)*<br>Department of Statistics<br>West Lafayette, IN 47907 | 7b. ADDRESS *(City, State, and ZIP Code)* |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>Office of Naval Research<br>8b. OFFICE SYMBOL *(If applicable)* | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER<br>N00014-88-K-0170, DMS-8606964, DMS-8702620 |
| 8c. ADDRESS *(City, State, and ZIP Code)*<br>Arlington, VA 22217-5000 | 10. SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO. |

11. TITLE *(Include Security Classification)*
ON A CLASSIFICATION PROBLEM: RANKING AND SELECTION APPROACH

12. PERSONAL AUTHOR(S)
Shanti S. Gupta and Lii-Yuh Leu

| 13a. TYPE OF REPORT | 13b. TIME COVERED | 14. DATE OF REPORT *(Year, Month, Day)* | 15. PAGE COUNT |
|---|---|---|---|
| Technical | FROM ______ TO ______ | August 1989 | 19 |

16. SUPPLEMENTARY NOTATION

| 17. COSATI CODES | | | 18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)* |
|---|---|---|---|
| FIELD | GROUP | SUB-GROUP | classification rules; multivariate normal populations; Mahalanobis distance; ranking and selection; probability of correct classification. |

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This paper deals with a classification problem based on ranking and selection approach. We assume that the populations follow multivariate normal distribution. The corresponding selection problem is to choose the population with the smallest Mahalanobis distance. The subset selection approach is considered throughout this paper. Sometimes the indifference zone approach is also proposed. It should be pointed out that, for the subset selection approach, we need not assume that the individual to be classified belongs to one of the several given categories. The classification procedures depend on whether the parameters $\underline{\mu}_i$ and $\Sigma_i$ are known or unknown.

| | |
|---|---|
| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS | 21. ABSTRACT SECURITY CLASSIFICATION<br>Unclassified |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>Shanti S. Gupta | 22b. TELEPHONE *(Include Area Code)*<br>(317) 494-6031 / 22c. OFFICE SYMBOL |

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED